**Data Science Project Protocol**

*Author:*

*Hagit Gadot*

# Introduction

The background hypothesis for this project is that **certain personal’s behavior and characteristic data are linked to the likelihood of using heavy drugs before the age of 24**.

A subset of this hypothesis was discussed as part of a [research](https://www.drugabuse.gov/news-events/nida-notes/2018/05/cigarette-smoking-increases-likelihood-drug-use-relapse) analysis published on the *National Institute on Drug Abuse* during May 2018 by Eric Sarlin, M.Ed., M.A. The analysis was based on data provided by 5,515 people who responded to the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) in 2001–2002 and again in 2004–2005.

The researchers findings were that cigarette smoking increased the likelihood of relapse among people in recovery from substance use disorder (SUD) and suggests that helping patients quit and remain abstinent from smoking may improve their chances for sustained recovery from use of other drugs. They have also found that people who smoked cigarettes at the initial interview and who were still smoking 3 years later were about 1.5 times more likely to use drugs and twice as likely to have SUD at follow-up than those who quit smoking. Among nonsmokers at the initial interview, those who had started to smoke between interviews were almost 5 times more likely to report substance use at the follow-up compared with those who did not smoke

The researchers cite possible explanations why cigarette smoking might increase the likelihood of SUD relapse to include that smoking often accompanies illicit drug use, and cigarettes may serve as a drug cue and relapse trigger in addition to linking nicotine exposure to cravings for stimulants and opiates.

The purpose of the present project is to further expand the analysis of the collected data and test the hypothesis as well as to study the strength and patterns of associations among these behaviors or characteristics. **The results will permit a more precise description of an individual profile to identify whether it is part of a groups at higher risk.**

# Methodology (Project design)

## Data

Using data for respondents in years 2011-2018 from the Third National Health and Nutrition Examination Survey (NHANES III) ([link](https://www.cdc.gov/nchs/index.htm) ).

1. Data tables:

* **Table 1: DUQ**

The Drug Use questionnaire. focuses on lifetime and current use of marijuana or hashish, cocaine, heroin, and methamphetamine, as well as intravenous use of these and other drugs.

Participants aged 12-69 years were eligible; however, only data from participants aged 18-69 are included in this file for public release

* **Data 2 :DMO**

The demographics file provides individual, family, and household-level information on the following topics:

• Survey participant’s household interview and examination status.

• Pregnancy, Household and family income and size, No. of children, and adults

• Other selected demographic information, such as gender, age, race/Hispanic origin, education, marital status, military service status, country of birth, citizenship, and years of U.S. residenc

* **Table 3: ECQ**

**I was unable to use the ECQ (Early Childhood) data** due to the protection confidentiality of survey participants

* **Table 4: HIQ**

The Health Insurance questionnaire provides types of insurance coverage

* **Table 5: SMQ**

The Smoking - Cigarette Use dataset provides a history of cigarette use, age at initiation, past 30-day use, were new in 2015-16 (18 years and older).

* **Table 6: HSQ**

The “Current Health Status” survey, provides personal interview data on overall health assessment.

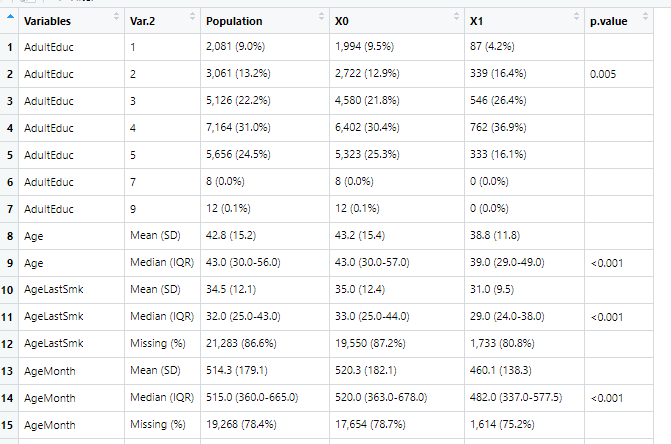
1. **The subject** – The characteristics of one who starts using ‘heavy’ drugs (i.e : cocaine/heroin/methamphetamine) before age 24.

The outcome is categorical (1=yes,0-no), Inclusion criteria- find the specific survey behavior and profile.

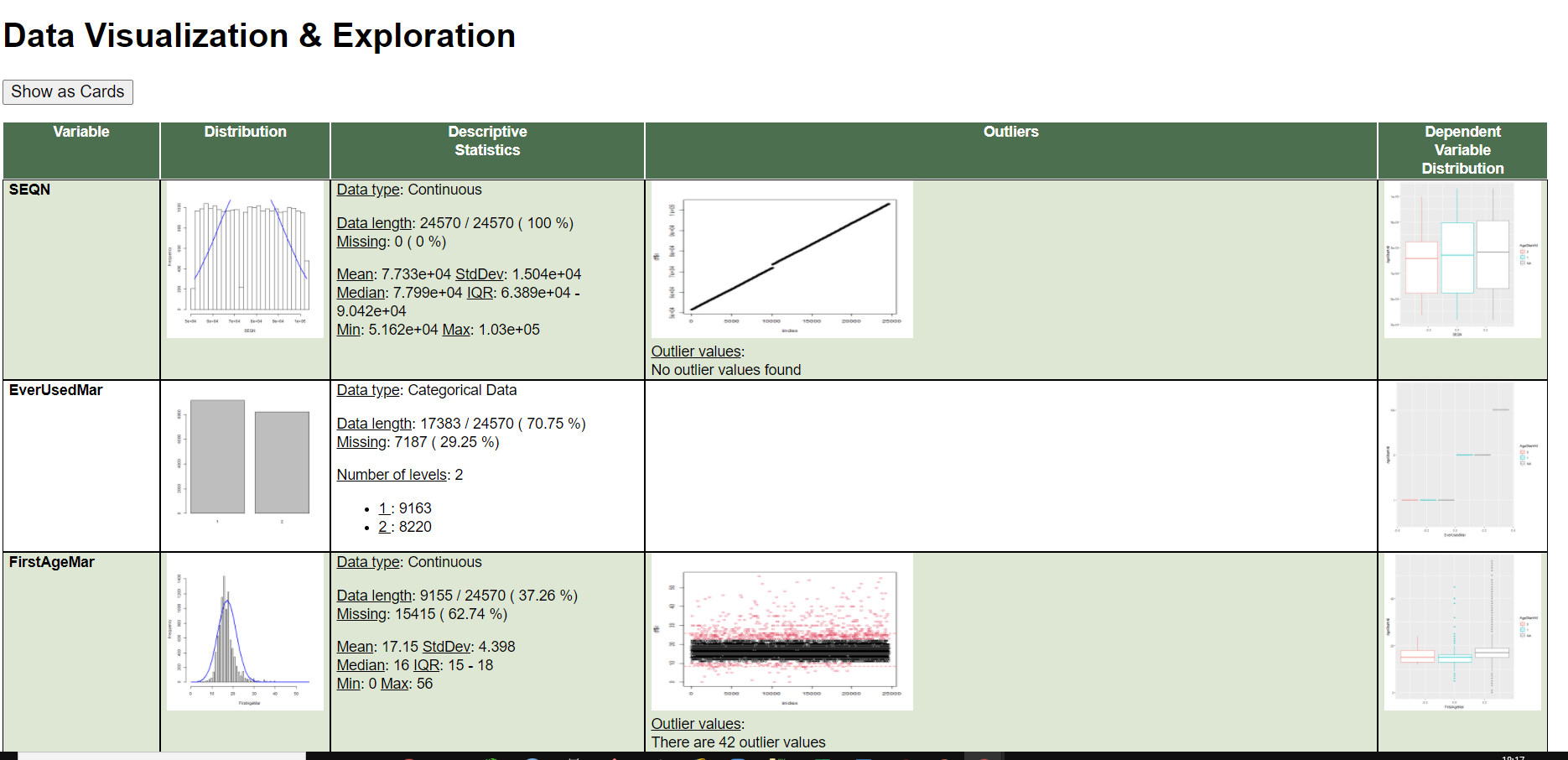
1. **Variables that may affect the outcome**: smoking behavior use, some demographic data such as household Income, army service, health insurance
2. **data exploration**

Using summary () command shows statistics base parameters (min, max,1,3 standard deviation, mean, median and NA count) for numeric variables, and unique value count for the categorical variable.

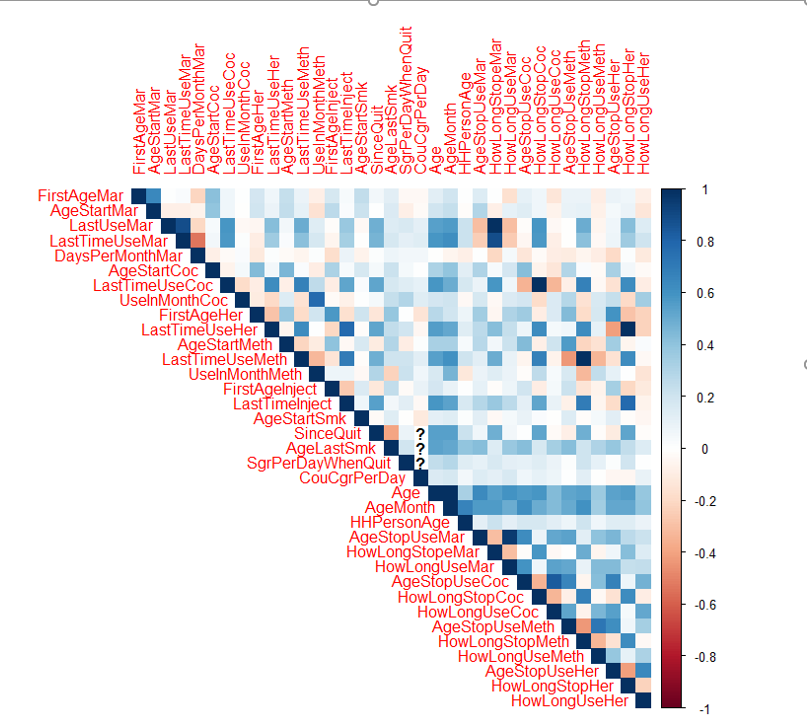
Using function table1()- shows the count, % of each variable category and calculates the Variance P-value between the variable and the outcome.



Using the ExploreData () shows histograms, boxplots, Descriptive statistics, outliers and Dependent Variable Distribution



Correlation using rcorr (), include Heat-map between all numeric columns



1. In data Enrichment phase I creates new features concludes intervals between given dates attributes that describe , for example how long the Participant used drugs or calculate the age first starts drugs
2. **Outliers**
   1. **Distribution change test**Using outliers matrix, I used Kolmogorov–Smirnov test – due to the fact that it is nonparametric methods for comparing two samples, sensitive to differences in both location and shape of the distribution functions of the two samples.
   2. **Correlation change test**using package ‘cocor’ - Statistical tests for the comparison between two correlations based on either independent or dependent groups.
   3. **Conclusion**If both correlation changed significant (p-val<0.05), andthe distribution changed also significantly – we leave the outliers (only one variable-“AgeStartSmoke”), all other combination- was replaced with NA.
3. **missing values**

**No Feature was removed because of the missing.**  the data of drug addiction is rarely report.

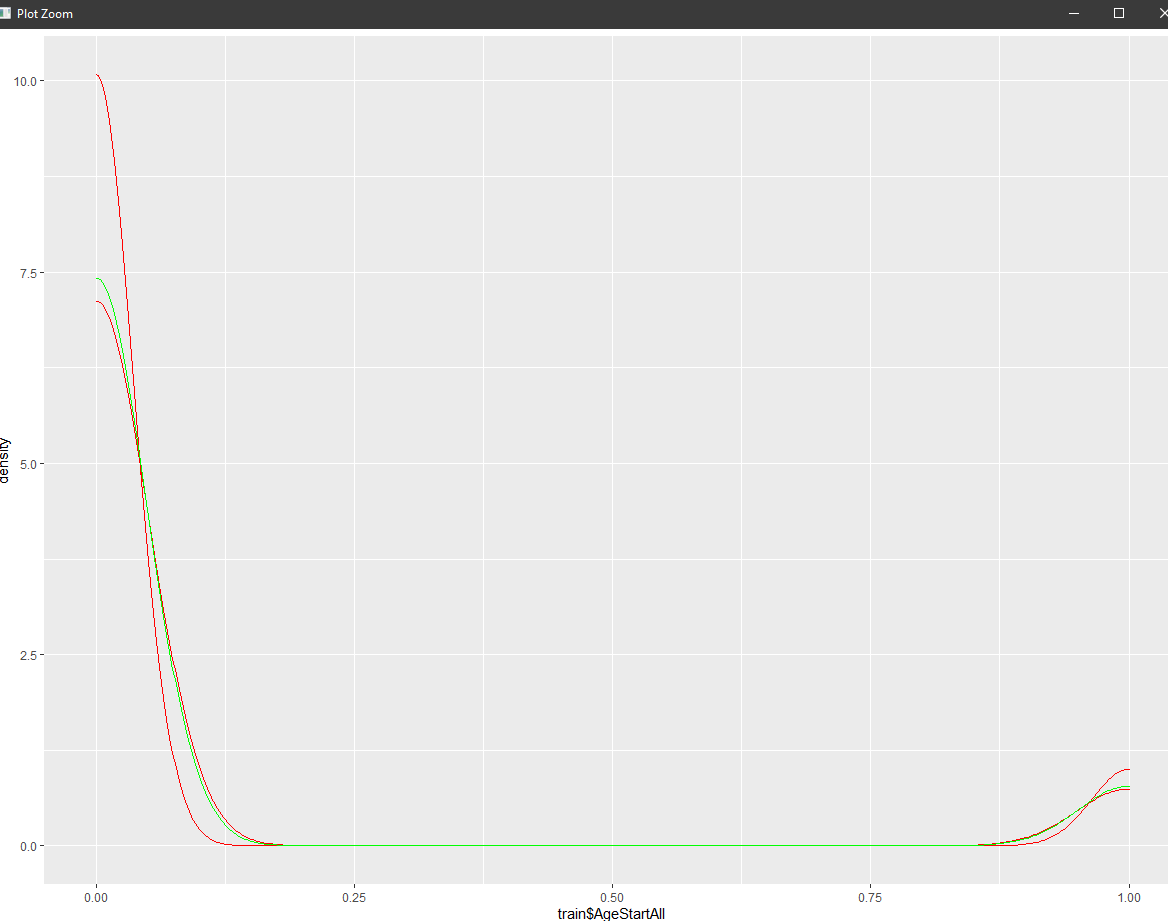
* 1. Removing more than 70% missing in a row- meaning delete 1025 rows (9%)
  2. All categoric variables – remain categoric – with NA value as 999.
  3. The numeric variable with more the 40% missing – turned to categoric variable classed due to how far the instance from the mean (in standard deviation count)
  4. The 3 numeric variables left had no outliers

1. Add at the end of the protocol (appendix) the [Data retrieval protocol](https://docs.google.com/spreadsheets/d/1pYYjgwZ_8PS1Bcmc2kRNHTL0f_rk__GCJALLs1JHPUQ/edit#gid=0)

## Models

Here you have to describe how do you plan to develop your models:

* The data was divided to train 80/20 the dev was also split in same ratio (80/20).  
  I used ‘train\_test()’ from ‘mechkar’ package. The split was perfectly balanced.



* The model outcome is classification type, so, Classification models were preformed:
  + k-nearest neighbor algorithm
  + Decision Tree
  + Support Vector Machines
  + Regression: Logistic Regression model
  + AdaBoost
  + Gradient Boosting Machine
  + Random Forest
* The measures that used to train and evaluate the models, were chosen by preforming feature selection vote using several analysis: univariant, multivariant , Random Forest, SVM….. the feature that got 2 and more votes, were taken as the model final features.
* The **Logistic Regression** model showed the best metric scores (AUC and Accuracy), it was also chosen because it is easier model to preform and explain.

## Deployment of your model

* The project files are export/ import from folder named: HagitGadot (../HagitGadot/filename.csv)
* Platforms to be deployed – demonstrate in a flow graph

**Files:**

train.csv

dev.csv

TheTest.csv

TheTrain.csv

**Files:**

VarSelect.scv

FeatureDF.csv

**Files:**

Selected.csv

* Feature selected:
  + '**CildEduc\_2' /'CildEduc\_8','CildEduc\_9'** 2’nd grade **/**8- 9 years at school
  + **'AgeStartMar\_10'** – start using Marijuana before age 10
  + **'FriqUseMar\_5'**- using Marijuana 24 times per month
  + **'LastUseMar\_25'**- before 17-21 years
  + **'Citizen\_1'**- Citizen by birth
  + **'FamilyIncome\_2'/'FamilyIncome\_3'**

Annual income of $ 5,000 to $ 14,999

* + **'Race\_2'**- Other Hispanic /**'Race\_3'** - Non-Hispanic White
  + **'CouCgrPerDay\_-15','CouCgrPerDay\_25'**- ,**'CouCgrPerDay\_30'**-  
    less than 5/ more the 25 cigarettes per day
  + **'YearsInUS\_1'** / **'YearsInUS\_2'/ 'YearsInUS\_9'** / **'YearsInUS\_8'**/ **'YearsInUS\_7'**  
    0-5 years in US or 30-50 years in US
  + ,**'AdultEduc\_1','AdultEduc\_6','AdultEduc\_7'**

Adult -Less then 9 years **or** more than collage

* + **'AgeStartMar\_-5','AgeStartMar\_-15','AgeStartMar\_30'**Age Start Marijuana 7-10 /older then 40
  + **'AgeStopUseMar\_-25'**- Age Stop use Marijuana- 38-43
  + **'Children6to17\_4'**- 3 or mor children
  + **'ColdLastMonth\_1','ColdLastMonth\_2'**- Had cold last month: yes and no
  + **'X.PeoInFamily\_1'**- 1 person in the family

# Results

#Here you will present the main results of all the processes. We will describe:

**Main data count:**

|  |  |  |  |
| --- | --- | --- | --- |
| Data type | Row | Columns | details |
|  |  |  |  |
| Input raw file | 24,570 | 67 | based on 30 files joined from NHANSE |
| Enriched data | 24,570 | 80 |  |
| Outliers data | Total out :837 |  |  |
| Missing in a row | Total miss:1025 |  | more of 70% in a row |
| Missing in a col | Total miss:7187 |  | Total 29.3% missing- categorization |
| Feature selected data | 23,545 | 38 |  |
| Flat numeric DF | 23,545 | 327 |  |
| Feature selected flat data | 23,545 | 267 |  |
| Feature selected data | 23,545 | 38 | Choose by voting |
| Train | 18,836 | 38 |  |
| Train2 | 15,068 | 38 |  |
| Dev | 3,769 | 38 |  |
| Test | 4,709 | 38 |  |

**Model prediction result (train/dev):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | model | accuracy-Train | accuracy-Test | AUC-train | AUC-test |
| 1 | Logistic Regression | 0.926268 | 0.916667 | 0.869323 | 0.845624 |
| 6 | AdaBoost | 0.926533 | 0.917197 | 0.868296 | 0.842874 |
| 7 | Gradient Boosting Machine | 0.929851 | 0.916401 | 0.872286 | 0.840587 |
| 3 | Random Forest | 0.910937 | 0.905520 | 0.861779 | 0.838919 |
| 2 | Decision Tree | 0.952150 | 0.902866 | 0.932911 | 0.674185 |
| 4 | SVM | 0.932174 | 0.919055 | 0.191240 | 0.281355 |
| 5 | kNN | 0.926998 | 0.901274 | 0.854752 | 0.235631 |

**Model Final prediction result (train/test):**

----------------------

accuracy-Train 0.92472

accuracy-Test 0.91994

AUC-train 0.86446

AUC-test 0.86064

# Conclusion

The purpose of the present project was to test the hypothesis as well as to study the strength and patterns of survey data among drug behaviors as presented in the sample.

The challenges were along all the project’s work: to determine the outcome, to deal with the NA’s in most of the data, programming in R and PYTHON.

The project is limited of prediction of future detection of drugs addiction.

1. The data did not include historical personal data (the youth data was restricted) and thus- was not include in the prediction.

2. The features that was assembled in the project were regarding behavior that my come after the age 24, but still could predict the hypothesis,

3. other features were answers related to the time of the survey (income, health insurance, ets), and may be relevant to the subject’s future.

4. In total point of view of the feature – the features were taking by the statistic impact on the outcome (the feature of the heavy frug use were filtered from the chosen features and not participate in the models.

The project can be use with the same file, but not in different data, that might predict wrong result, or even not work at all (no prediction).

In the program there were several places where the conclusion was manually – such as feature to select (how many votes co include), which model to select ….

Therefor, the project is not generic and not “stand alone”.

**Personal note:**

This course was very different from “class course”, I decided to share two thoughts over the course:

* The recorded lessons were fantastic! I wish all my study had been like that!
* I think there was not enough supervised practice during the course .  
   I felt it when I needed to do it on my own.

Thanks for being patient along the whole course.